Econometric analysis of vast covariance matrices using composite realized kernels*

ASGER LUNDE
School of Economics and Management, University of Aarhus,
Bartholins Allé 10, 8000 Aarhus C, Denmark
& CREATES, University of Aarhus
alunde@econ.au.dk

NEIL SHEPHARD
Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
& Department of Economics, University of Oxford
neil.shephard@economics.ox.ac.uk

KEVIN SHEPPARD
Oxford-Man Institute, University of Oxford,
Eagle House, Walton Well Road, Oxford OX2 6ED, UK
& Department of Economics, University of Oxford
kevin.sheppard@economics.ox.ac.uk

August 25, 2011

Abstract

We propose a composite realized kernel to estimate the ex-post covariation of asset prices. Composite realized kernels are a data efficient method where the covariance estimate is composed of univariate realized kernels to estimate variances and bivariate realized kernels to estimate correlations. We analyze the merits of our composite realized kernels in an ultra high dimensional environment, making economic decisions every day solely based on the previous day's data. The first application is a minimum variance portfolio exercise and this is followed by an investigation of portfolio tracking. The data set is tick-by-tick data comprising 473 US equities over the sample period 2006-2009. We show that our estimator is able to deliver a significantly lower portfolio variance than its competitors.

Keywords: Covariance; High frequency data; Market frictions; Minimum variance portfolio; Realized kernel; Tracking portfolio.

*We thank ... for helpful comments on this paper.

1 Introduction

1.1 Background to this paper

Multivariate volatility measurement and forecasting has becoming increasingly important not only because of its direct application in portfolio allocation and asset pricing, but also due to the insights they provide into risk management practices (e.g. using low-frequency data, Brownlees & Engle (2010) portray the importance of modelling conditional correlations for systemic risk management, where they show that a rise in a firm's stock volatility and correlation with the market magnifies its contribution to their proposed measure of systemic risk.)

The econometric analysis of multivariate financial data is challenging because of both the usual univariate market frictions and the additional feature of non-synchronous trading.

Non-synchronous trading leads common covariance statistics to have a severe bias towards zero. This phenomenon is often referred to as the Epps effect, as Eps (1979) found this bias for stock returns. This effect is frustrating for econometric theory suggests that high frequency data should potentially yield significant information about the comovements of financial assets. The Eps effect limited the early exploitation of high frequency data using realised covariances formalised by Andersen et al. (2003) and Barndorff-Nielsen & Shephard (2004). Their work had to be applied using sparse sampling of returns of 5-20 minutes to try to mitigate the impact of noise and non-synchronous trading. In several recent studies it has been confirmed that realised covariances computed over shorter fixed time periods underestimate the degree of dependence between assets (See e.g. Renò (2003), Hayashi & Yoshida (2005), Voev & Lunde (2007), Griffin & Oomen (2011), Zhang (2011) and Barndorff-Nielsen et al. (2011)).

The multivariate realised kernel suggested by Barndorff-Nielsen et al. (2011) utilize refresh time sampling to synchronize the timing of a multivariate data set and employ leads and lagged autocovariance terms to mop up the remaining effects of the noise and non-sychroneity. This combined with an appropriate weight function allows the construction of a consistent positive semi-definite estimator. The desireable feature of positive semidefiniteness comes with a cost of important data loss in high dimensions. It is shown in Barndorff-Nielsen et al. (2011) that the multivariate realised kernel works well in moderate dimensions (≤ 30) when all the assets are very frequently traded. It is however also a fact that the refresh time sampling scheme will be controlled by the least frequently traded asset.

Hautsch et al. (2011) addressed the sensitivity of the realised kernels to the least frequently traded asset by developing a blocking strategy that separate groups of more liquid assets from groups of less liquid asset. Then they apply the multivariate realized kernel to each group and recombine the parts in a clever way. Their resulting estimator is not positive semi-definite, a problem which they handled using a method called eigenvalue cleaning we will explain later.

1.2 Contribution of this paper

The contribution of this paper is to introduce the *Composite Realized Kernel*. This estimates each entry of the integrated covariance matrix optimally in the sense that each entry will

- have its bandwidth tailor made to glean efficiency,
- utilizes all available data points, and
- has its weight function selected to generate efficiency.

This approach is inspired by, but is distinct from, the composite likelihood literature. This discusses estimating large dimensional models by looking at many small dimensional submodels. In our case we are interested in large dimensional covariances and we will carry this out by seperately estimating the individual variances and the individual correlations optimally. Of course our approach is not likelihood based, but the composite theme is still strong. The origins of the composite likelihood method go back to at least Lindsay (1988). See Varin (2008) and Varin et al. (2011) for a review.

1.3 Corresponding literatures

1.3.1 Measuring multivariate financial variation

Taking a step back for a moment, the use of high-frequency inspired realised measures, based around the concept of quadratic variation, was first formalised in financial econometrics by Andersen et al. (2001) and Barndorff-Nielsen & Shephard (2002) in their work on univariate realised variances. Extensive work has developed which formally allowed for the effect of univariate market microstructure effects on the analysis (e.g. Zhang et al. (2005), Barndorff-Nielsen et al. (2008) and Jacod et al. (2009), Zhang (2006) and Xiu (2010)), others use moderate frequency data such as five minute returns in order to ameliorate the effect (e.g. Andersen et al. (2000) and Bandi & Russell (2008)). The composite realised kernel directly applies this research in its estimation of the individual diagonal elements of the covariance matrix.

In the multivariate case much less work has been carried out. Important contributions include work dealing with non-synchronous data include Bannuh et al. (2009), Clément & Gloter (2011), Hayashi & Yoshida (2005), Malliavin & Mancino (2002, 2009) and Wang & Zou (2010). Allowing for noise as well as non-synchronous trading is the focus of Zhang (2011), Christensen et al. (2010), Aït-Sahalia et al. (2010) and Barndorff-Nielsen et al. (2011). The latter paper is our focus as it is the only one we know which is robust and delivers positive semi-definite estimators, and so estimates of correlations in a way which obeys trivial bounds. We change the approach in this paper though, as our focus will be in estimating individual correlations, as contributions to the composite realised kernel matrix.

1.3.2 Forecasting multivariate financial variation

Composite realised kernels estimate the quadratic variation of a vector of financial assets, so measuring how variable and dependent they have been in the past. To make economic decisions we have to forecast. There has been a significant literature taking realised measures and extrapolating them into the future. In this paper we will use a simple random walk forecasting device, but we should take a moment to put this in the context of the more sophisticated forecasting literature on this topic.

There is some recent research that focuses only on modelling and forecasting the realized covariance matrix; see, for example, Voev (2008), Chiriac & Voev (2011) and Bauer & Vorkink (2011). The focus in these studies is on developing parsimonious models to forecast the realized covariance matrix. In contrast, Noureldin et al. (2011) develops a framework for forecasting the covariance of daily returns which also requires forecasts of the realized measure. They find the realized measure to be a more precise factor to drive the volatility dynamics for daily returns compared to the outer product of daily returns which is used in GARCH models.

Jin & Maheu (2010) utilizing realized measures to improve the density forecasts of multivariate daily returns. Search is still ongoing for multivariate volatility models with flexible dynamics and ease of application in moderately large dimensions. The crisis forcefully demonstrated the need for more robust models to capture and project financial risk; in particular to capture correlation dynamics. However in practice, developing new models faces the "curse of dimensionality" in reference to the - often exponential - increase in the number of model parameters as the number of assets under study grows. Reviews of the multivariate generalised autoregressive conditional heteroskedasticity (GARCH) literature are given by, for example, Bauwens et al. (2006), Engle (2009), (Francq & Zakoian 2010, Ch. 11) and Silvennoinen & Teräsvirta (2009).

1.4 Outline of paper

The plan of the paper is as follows. In Section 2 we present the formal framework that will guide our estimation strategy. Building on the core results for the multivariate realized kernel we define our composite realized kernel. Because this estimator is not guaranteed to be positive semi-definite we present two ways of projecting it onto the space of positive defininte matrices. The first method is known as eigenvalue cleaning and the second imposes a factor structure on returns. Section 3 presents the data set and gives a descriptive analysis of the estimated high dimensional covariance and correlation matrices. In Section 4 we consider some applications in portfolio construction, sometimes imposing shortsale constraints. We analyse the merits of our composite realized kernels in an ultra high dimensional environment. The first application is a minimum variance portfolio exercise and this is followed by an investigation of portfolio tracking. Section 5 concludes the paper.

2 Covariance Estimation using High Frequency Data

Before we turn to the presentation of our estimators we need some notation and underlying assumptions. The framework follows Barndorff-Nielsen et al. (2011) so we study a $d-$dimensional log price process $X = \left(X^{(1)}, X^{(2)}, ..., X^{(d)}\right)'$. These prices are observed irregularly and non-synchronous over a generic interval $[0, 1]$, which we think of as a day.

We write the observation times for the i-th asset as $t_1^{(i)}, t_2^{(i)}, \ldots$, which could correspond to trades or quote updates. Hence, the database of prices at hand is $X^{(i)}(t_j^{(i)})$, for $j = 1, 2, ..., N^{(i)}(1)$, and $i = 1, 2, ..., d$. Here $N^{(i)}(t)$ counts the number of distinct data points available for the i-th asset up to time t.

The observed price process, X, is assumed to be driven by Y, the efficient price, abstracting from market microstructure effects. The efficient price is modeled as a *Brownian semimartingale* ($Y \in \mathcal{BSM}$)

$$Y(t) = \int_0^t a(u)\mathrm{d}u + \int_0^t \sigma(u)\mathrm{d}W(u),$$

where a is a vector of elements which are predictable locally bounded drifts, σ is a càdlàg volatility matrix process and W is a vector of independent standard Brownian motions. If $Y \in \mathcal{BSM}$ then its ex-post covariation is given by

$$[Y](1) = \int_0^1 \Sigma(u)\mathrm{d}u, \quad \text{where} \quad \Sigma = \sigma\sigma',$$

so in this model the quadratic variation of Y, $[Y](1)$, equals the integrated covariance matrix which is the object of econometric interest.

2.1 Realized Kernel and Bandwidth Selection

Our main focus in this paper is repeatedly using the multivariate realized kernel (MRK) that was suggested by Barndorff-Nielsen et al. (2011). Their contribution is to construct a consistent, positive semi-definite (psd) estimator of $[Y](1)$ from the available database of asset prices. The construction deals with the three major challenges of high frequency financial data:

1. there are market microstructure effects $U = X - Y$,
2. the data is irregularly spaced and non-synchronous,
3. the market microstructure effects are not statistically independent of the Y process.

To deal with the non-synchronicity of the data we apply the *refresh time* definition of Barndorff-Nielsen et al. (2011)[1].

[1]Refresh time was used in a cointegration study of price discovery by Harris, McInish, Shoesmith & Wood (1995). Martens (2003) used the same idea in the context of realised covariances, but his estimator is inconsistent.

2.1.1 Preprocessing the data

Definition 1 *We define the first refresh time as* $\tau_1 = \max\left(t_1^{(1)}, ..., t_1^{(d)}\right)$*, and then subsequent refresh times as*

$$\tau_{j+1} = \max\left(t^{(1)}_{N^{(1)}_{\tau_j}+1}, ..., t^{(d)}_{N^{(d)}_{\tau_j}+1}\right).$$

The resulting Refresh Time sample size is N*, while we write* $n^{(i)} = N^{(i)}(1)$*.*

So τ_1 is the time it has taken until all the assets has traded, i.e. all the posted prices have been updated. τ_2 is the first time when all the prices are again refreshed.

Note that N is random and we write the durations as $\tau_{N,i} - \tau_{N,i-1} = \Delta_{N,i} = \frac{D_{N,i}}{N}$ for all i. We will however omit the subscript-N and make the dependence on N implicit. We will follow Barndorff-Nielsen et al. (2011) and work under the following assumptions about the durations between observation times.

Assumption D. (i) *That* $\mathrm{E}(D^r_{\lfloor tN \rfloor} | \mathcal{F}_{\tau_{\lfloor tN \rfloor - 1}}) \xrightarrow{p} \varkappa_r(t)$, $0 < r \leq 2$, *as* $N \to \infty$. *Here we assume* $\varkappa_r(t)$ *are strictly positive càdlàg processes adapted to* $\{\mathcal{F}_t\}$; (ii) $\max_{i \in \{j+1,...,j+R\}} D_i = o_p(R^{1/2})$ for any j; (iii) $\tau_0 \leq 0$ *and* $\tau_{N+1} \geq 1$.

2.1.2 Constructing a multivariate realised kernel

Having defined the common time clock, $\{\tau_j\}$, we can now construct the vector returns series that the multivariate realized kernel will be based on. Let $n, m \in \mathbb{N}$, with $n - 1 + 2m = N$, and define the vector observations $X_0, X_1, ..., X_n$ as $X_j = X(\tau_{j+m})$, $j = 1, 2, ..., n-1$, and

$$X_0 = \frac{1}{m}\sum_{j=1}^{m} X\left(\tau_j\right) \quad \text{and} \quad X_n = \frac{1}{m}\sum_{j=1}^{m} X(\tau_{N-m+j}).$$

So X_0 and X_n are constructed by jittering initial and final time points.[2] We can now define the high frequency vector returns:

$$x_j = X_j - X_{j-1}, j = 1, 2, ..., n.$$

The class of positive semi-definite *multivariate realised kernels* (RK) has on the following form

$$K(X) = \sum_{h=-n}^{n} k(\tfrac{h}{H})\Gamma_h, \qquad \text{where } \Gamma_h = \sum_{j=h+1}^{n} x_j x'_{j-h}, \text{ for } h \geq 0, \tag{1}$$

and $\Gamma_h = \Gamma'_{-h}$ for $h < 0$ is the h-th realised autocovariance.

Following Barndorff-Nielsen et al. (2011), the non-stochastic weight function, $k : \mathbb{R} \curvearrowright \mathbb{R}$, will be taken to be of Parzen form. In particular this means that

[2] For details about jittering see Barndorff-Nielsen et al. (2011).

$$k(x) = \begin{cases} 1 - 6x^2 + 6x^3, & 0 \leq x \leq 1/2, \\ 2(1-x)^3, & 1/2 \leq x \leq 1, \\ 0, & x > 1. \end{cases}$$

2.1.3 Some assumptions

The parzen form satisfies the Barndorff-Nielsen et al. (2011) assumption

Assumption K. (i) $k(0) = 1, k'(0) = 0$; (ii) k is twice differentiable with continuous derivatives; (iii) define $k_\bullet^{0,0} = \int_0^\infty k(x)^2 \mathrm{d}x$, $k_\bullet^{1,1} = \int_0^\infty k'(x)^2 \mathrm{d}x$, and $k_\bullet^{2,2} = \int_0^\infty k''(x)^2 \mathrm{d}x$ then $k_\bullet^{0,0}, k_\bullet^{1,1}, k_\bullet^{2,2} < \infty$; (iv) $\int_{-\infty}^\infty k(x) \exp(ix\lambda) \mathrm{d}x \geq 0$ for all $\lambda \in \mathbb{R}$.

The Parzen form of k means that $|k''(0)| = 12$ and $k_\bullet^{0,0} = 0.269$. The implications of other choices for k are discussed in Barndorff-Nielsen et al. (2008, 2011).

So the multivariate realised kernel has a similar form as a standard heteroskedasticity and autocorrelated (HAC) covariance matrix estimator familiar in econometrics (e.g. Newey & West (1987) and Andrews (1991)). However, no adjustment is made to take out the mean, the data we use are temporal differences, not levels, and we have to be careful about end conditions.

Finally, we must write out our assumptions about the market microstructure effects U that govern the properties of the vector returns $\{x_j\}$ and so $K(X)$. We define the noise associated with $X(\tau_{N,j})$ at the observation time $\tau_{N,j}$ as $U_{N,j} = X(\tau_{N,j}) - Y(\tau_{N,j})$. We work under Assumption **U** in Barndorff-Nielsen et al. (2011), but it suffices to state that we denote the average long run variance of U by Ω that is a $d \times d$ matrix. Moreover, we denote by SH the assumption that a and σ are bounded.

2.1.4 Sampling behaviour

With all these details in place we have the core theorem on which all our subsequent estimators are built. This is stated using a matrix normal notation. Our result will use the matrix normal distribution. For $M \in R^{q \times q}$, $M \sim N(A, B)$ simply means that $vec(M)$ is Gaussian distributed with mean $vec(A)$ and the covariance between $a'Mb$ and $c'Md$ is given by $cov(a'Mb, c'Md) = v_{ab} B v_{cd}$, with $v_{ab} = vec(\frac{ab' + ba'}{2})$ and $v_{cd} = vec(\frac{cd' + dc'}{2})$.

Theorem 1 *Suppose assumptions D, U, SH and K in Barndorff-Nielsen et al. (2011) holds. Let* $H = c_0 n^{3/5}$, *then*

$$n^{1/5} \left\{ K(X) - \int_0^1 \Sigma(u) \mathrm{d}u \right\} \xrightarrow{Ls} \mathrm{MN}\{c_0^{-2} |k''(0)| \Omega, 4 c_0 k_\bullet^{0,0} \Psi\}.$$

Here

$$\Psi = \int_0^1 \{\Sigma(u) \otimes \Sigma(u)\} \frac{\varkappa_2(u)}{\varkappa_1(u)} \mathrm{d}u$$

is the $d^2 \times d^2$ random matrix that extend the definition of integrated quarticity to the multivariate context, with $\otimes$ denoting the Kronecker product.

A main feature of multivariate kernels is that there is a single bandwidth parameter H which controls the number of leads and lags used for all the series. Our benchmark approach will be to follow Barndorff-Nielsen et al. (2011) and first apply the univariate optimal mean square error bandwidth selection to each asset price. This gives the d individual bandwidths and the ad hoc method over averaging is used to choose the global H, simply by

$$\bar{H} = d^{-1} \sum_{i=1}^{d} H^{(i)}.$$

It has been noted by for example Barndorff-Nielsen et al. (2011) and Hautsch et al. (2011) that the realized kernel estimator can be sensitive to the different sampling frequencies by which we observe the individual assets. The estimation precision will effectively be dictated by the least liquid asset. To address this problem Hautsch et al. (2011) introduce a blocking strategy that separate groups of more liquid assets from groups of less liquid asset. Then they apply the multivariate realized kernel to each group with the corresponding bandwidth $\bar{H}$ for that group.

2.2 The Composite Realized Kernel

The composite realized kernel that we suggest takes a distinct approach. We will estimate each entry of the integrated covariance matrix optimally in the sense that each entry will have its bandwidth tailor made and the data loss will be minimal. We break now ground by fine tuning the estimation of each coordinate. For the diagonal elements, the integrated variances, the bandwidth is determined as in Barndorff-Nielsen et al. (2009). So for the i-th series the optimal bandwidth is $H_i = c_0 \left(n^{(i)}\right)^{3/5}$ with

$$c_0 = c^* \left\{ \frac{\omega_i^2}{\sqrt{\int_0^1 \sigma_i^4(u) du}} \right\}^{2/5}. \quad c^* = \left\{ \frac{k''(0)^2}{k_\bullet^{0,0}} \right\}^{1/5},$$

where $c^* \simeq 3.5134$ because we have chosen the Parzen kernel. Further, $\omega_i^2 = \Omega_{ii}$, the ii-th element of Ω.

In the composite realised kernel all the off-diagonal elements are determined through the estimation of a bivariate model, allowing us to determine an estimate of the realised correlation. In the bivariate case we can write Theorem 1 as a central limit theorm for the realized kernel for the variance of $X^{(i)}$, for the covariance of $X^{(i)}$ and $X^{(j)}$, and the variance of $X^{(j)}$

$$n^{1/5} \begin{pmatrix} K(X^{(i)}) - \int_0^1 \Sigma_{ii} \mathrm{d}u \\ K(X^{(i)}, X^{(j)}) - \int_0^1 \Sigma_{ij} \mathrm{d}u \\ K(X^{(j)}) - \int_0^1 \Sigma_{jj} \mathrm{d}u \end{pmatrix} \xrightarrow{Ls} \mathrm{MN}\left(c_0^{-2} |k''(0)| A, 2c_0 k_\bullet^{0,0} B\right), \tag{2}$$

where

$$A = \begin{pmatrix} \Omega_{ii} \\ \Omega_{ij} \\ \Omega_{jj} \end{pmatrix} \quad \text{and} \quad B = \int_0^1 \begin{pmatrix} 2\Sigma_{ii}^2 & 2\Sigma_{ii}\Sigma_{ij} & 2\Sigma_{ij}^2 \\ \bullet & \Sigma_{ii}\Sigma_{jj} + \Sigma_{ij}^2 & 2\Sigma_{jj}\Sigma_{ji} \\ \bullet & \bullet & 2\Sigma_{jj}^2 \end{pmatrix} \frac{\varkappa_2}{\varkappa_1} \mathrm{d}u.$$

Then by the delta method we can deduce the asymptotic distribution of the correlation

$$n^{1/5} \left(\frac{K(X^{(i)}, X^{(j)})}{\sqrt{K(X^{(i)})K(X^{(j)})}} - \rho_{ij} \right) \xrightarrow{Ls} \mathrm{MN}\left(c_0^{-2} |k''(0)| \omega_{cor(i,j)}, 2c_0 k_\bullet^{0,0} \mathrm{IQ}_{cor(i,j)} \right),$$

where

$$\omega_{cor(i,j)} = \frac{v'Av}{\sqrt{\int_0^1 \Sigma_{jj}\mathrm{d}u \int_0^1 \Sigma_{ii}\mathrm{d}u}}, \quad \mathrm{IQ}_{cor(i,j)} = \frac{v'Bv}{\left(\int_0^1 \Sigma_{ii}\mathrm{d}u\right)\left(\int_0^1 \Sigma_{jj}\mathrm{d}u\right)},$$

with

$$v = \begin{pmatrix} -\frac{1}{2}\beta_{ji} \\ 1 \\ -\frac{1}{2}\beta_{ij} \end{pmatrix}, \quad \beta_{ij} = \frac{\int_0^1 \Sigma_{ij}\mathrm{d}u}{\int_0^1 \Sigma_{jj}\mathrm{d}u}.$$

Moreover, the mean square error optimal bandwidth is $H = c_0 n^{3/5}$ with

$$c_0 = c^* \xi_{cor(i,j)}^{1/5}, \quad \xi_{cor(i,j)} = \frac{\omega_{cor(i,j)}^2}{\mathrm{IQ}_{cor(i,j)}}, \quad c^* = \left\{ \frac{k''(0)^2}{k_\bullet^{0,0}} \right\}^{1/5},$$

where again $c^* \simeq 3.5134$ as in the univariate case.

2.3 Implementing the realized kernels

To implement this we need plug in estimates of Ω and B. To estimate the covariance matrix of the noise, Ω, we use a simple HAC estimator with one lag only

$$\hat{\Omega} = \frac{\sum_{j=1}^n x_j x_j' + \frac{1}{2}\sum_{j=2}^n x_j x_{j-1}'}{2n}.$$

This correspond to a multivariate version of the univariate estimator advocated by Zhang et al. (2005) and Bandi & Russell (2008), but modified to take into account the refresh time sampling scheme.

For the quarticity part we pretend that Σ is time-invariant and so use a realized kernel to estimate B in the obvious way. That is IQ is estimated as RK squared, that is an estimator of the lower bound of IQ. So we use

$$\hat{B} = \begin{pmatrix} 2\widetilde{RK}_{ii}^2 & 2\widetilde{RK}_{ii}\widetilde{RK}_{ij} & 2\widetilde{RK}_{ij}^2 \\ \bullet & \widetilde{RK}_{ii}\widetilde{RK}_{jj} + \widetilde{RK}_{ij}^2 & 2\widetilde{RK}_{jj}\widetilde{RK}_{ij} \\ \bullet & \bullet & 2\widetilde{RK}_{jj}^2 \end{pmatrix}$$

where $[\widetilde{RK}_{ij}]$ is the 2×2 realized kernel optimized for variances. Here we could obviously do a second step and re-estimate using a bandwidth based on the first step *correlation optimal* kernel.

Having estimated the correlations optimally we proceed to variances. These are simply estimated using the non-negative univariate realized kernels implemented as in Barndorff-Nielsen et al. (2009). So let the vector

$$K_{RVol} = \left(\sqrt{K\left(X_1\right)} \quad , \sqrt{K\left(X_2\right)} \quad , \cdots \quad , \sqrt{K\left(X_d\right)} \right)',$$

which collects the univariate volatility estimates and let

$$K_{Cor} = \begin{pmatrix} 1 & \frac{K_{12}(X_1,X_2)}{\sqrt{K_{22}(X_1,X_2)K_{22}(X_1,X_2)}} & \cdots & \frac{K_{12}(X_1,X_d)}{\sqrt{K_{22}(X_1,X_d)K_{22}(X_1,X_d)}} \\ \bullet & 1 & \ddots & \vdots \\ \bullet & \bullet & \ddots & \frac{K_{12}(X_{d-1},X_d)}{\sqrt{K_{22}(X_{d-1},X_d)K_{22}(X_{d-1},X_d)}} \\ \bullet & \bullet & \bullet & 1 \end{pmatrix},$$

the composite realized correlation matrix. We construct the composite realized kernel as

$$K_{Cov} = K_{RVol} K'_{RVol} \odot K_{Cor},$$

where $\odot$ denotes the Hadamard product. Note that all the 2×2 parts used to construct K_{Cor} are positive semi-definite, and so $-1 \leq [K_{Cor}]_{ij} \leq 1$. Additionally $K\left(X_i\right)$ is also strictly positive by construction. However, K_{Cov} and K_{Cor} not necessarily positive semi-definite by construction, although they utilize all data points and have all components estimated optimally in an element-by-element sense.

2.4 Projection

Realized covariance and multivariate realized kernels are only guaranteed to be positive semi-definite, and in large panels will usually have many 0 eigenvalues. The composite realized kernel is also not ensured to be positive semi-definite, and in practice has negative eigenvalues. Positive definiteness of covariance estimators is a desirable feature, even when it is not essential. We consider two methods to transform a positive semi-definite or indefinite covariance to be positive definite.

2.4.1 Eigenvalue cleaning

The first is the regularization method used in Hautsch et al. (2011), which is known as eigenvalue cleaning or eigenvalue clipping. This form of regularization is implemented using the correlation matrix associated with an estimated covariance matrix, $\hat{\Sigma}$,

$$\hat{R} \quad \equiv \quad \left(\hat{\Sigma} \odot I\right)^{-\frac{1}{2}} \hat{\Sigma} \left(\hat{\Sigma} \odot I\right)^{-\frac{1}{2}}$$

using the singular value decomposition $\hat{R} = \hat{P}\hat{\Lambda}\hat{P}'$ where $\hat{\Lambda}$ is a diagonal matrix containing the eigenvalues and $\hat{P}$ is the orthonormal matrix of associated eigenvectors. We assume the eigenvalues are ordered from largest to smallest so that $\lambda_i \geq \lambda_j$ whenever $i \geq j$. The eigenvalue cleaning is implemented by averaging all eigenvalues below the threshold $\lambda_{\max} = \left(1 + d/N + 2\sqrt{d/N}\right)$ where N is the number of returns used to estimate the covariance. This choice comes from Random Matrix Theory and is discussed in some detail in the Appendix. The first use of this approach we have seen in finance is due to Laloux et al. (2000). Eigenvalues less than $\lambda_{\max}$ are replaced by the average value of the eigenvalue less than this threshold,

$$\tilde{\lambda}_i = \frac{\sum_{j=k}^{d} \max\left(0, \lambda_j\right)}{d-k}, \; i = k, \ldots, d$$

where k is the index of the largest eigenvalue smaller than $\lambda_{\max}$.

The regularized covariance is then constructed using the original eigenvalues in positions $1, \ldots, k-1$ and $\tilde{\lambda}_i$ in the remaining positions,

$$\tilde{R} = \hat{P}\tilde{\Lambda}\hat{P}.$$

In practice Hautsch et al. (2011) recommend tightening the threshold by using

$$\lambda_{\max} = (1 - \lambda_1/d)\left(1 + d/N + 2\sqrt{d/N}\right)$$

which allows for the detection of more non-noise eigenvalues and reflects that the correlation structure of returns often has a large, dominant factor. The regularized covariance is then constructed as $\left(\hat{\Sigma} \odot I\right)^{\frac{1}{2}} \tilde{R} \left(\hat{\Sigma} \odot I\right)^{\frac{1}{2}}$. When used with realized covariance or the multivariate kernel computed on all assets, N is the number of returns used: 78 in the case of the realized covariance and the number of refresh time returns in the case of the multivariate kernel. When using applying eigenvalue cleaning to composite realized kernels, the number of returns used differs across the elements of the covariance matrix, and we used the minimum number of refresh times across all bivariate pairs for N.

2.4.2 Factor models

The second method exploits the factor structure of returns using 1- and 3-factor approximations. Factor approximation were shown to have fast convergence for estimating the inverse of a covariance matrix in Fan, Fan & Lv (2008). They are also popular in the finance literature (Chan, Karceski & Lakonishok 1999, Briner & Connor 2008) where imposing a factor structure is often found to improve covariance forecasts in portfolio optimization. Using the SVD on the correlation,

$$\hat{R} = \hat{P}\hat{\Lambda}\hat{P}',$$

the factor based projection is constructed using the largest q eigenvalues and their associated eigenvectors. Assume the that eigenvalues are ordered from largest to smallest so that the

1st diagonal position of $\hat{\Lambda}$contains the largest eigenvalue. The factor-based covariance is constructed using the upper q by q block of $\hat{\Lambda}$, denoted $\tilde{\Lambda}$, and the first q columns of $\hat{P}$, denoted $\tilde{P}$,

$$\tilde{R} = \tilde{P}\tilde{\Lambda}\tilde{P}' + \left(I - \left(\tilde{P}\tilde{\Lambda}\tilde{P}'\right) \odot I\right).$$

The diagonal elements of $\tilde{R}$ are identical to those of $\hat{R}$ and the correlation are determined by the factor structure. In the application we considered $q = 1$ and 3.

3 Data

We analyze high-frequency assets prices for a large selection of large US companies. Our base sample comprised about 600 stocks that appeared in the SP500 at some time in the sample period January 2006 though December 2009. We restricted attention to exclude rather illiquid stocks in the sense that we require a stock to trade a least 195 times a day to remain in the sample. This left us with 473 stocks that we followed over 996 trading days.

The data is the collection of recorded trades taken from the TAQ database accessed through the Wharton Research Data Services (WRDS) system. We followed the step-by-step cleaning procedure used in Barndorff-Nielsen et al. (2009) who discuss in detail the various choices available and their impact on univariate realized kernels. These cleaning rules include a suggestion that only data from a single exchange is used. We follow this by using the exchange for each asset on each day which had the largest number of trades. It should also be noted that we did not use any quote information in the data cleaning.

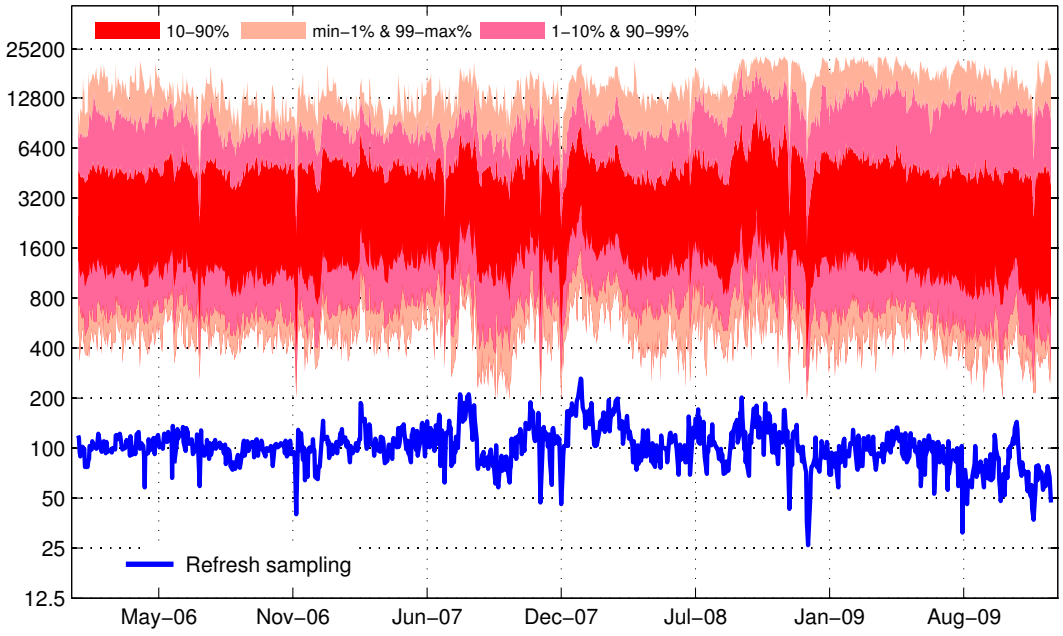

Figure 1: Daily observation count distribution for individual assets together with the time series of refresh time observations for the multivariate kernel using all assets [NOTE! Should we add min refresh counts across all bivariate pairs? ASGER: YES. PERHAPS A SEPERATE PLOT FOCUSING ON REFRESH TIME, IE FIG 1 IS ABOUT ESTIMATING VARIANCES AND FIG 1B ABOUT ESTIMATING CORRELATIONS?].

There are a substantial amount of variation in the sampling frequencies for the individual assets. In Figure 1 we show how the distribution of the number of observations evolve over the sample period. It is clear that the stocks in our sample are heavily traded, with around 90 per cent having more than thousand transaction a day.[3] Also included in Figure 1 is the time

[3]TAQ timestamps are only recorded with a resolution of 1 second, so we aggregated all trades with the same

series of refresh times based on all the assets. The refresh time sampling scheme, when prices are allowed to be stale by at most one tick, gives about 100 observations a day. It is interesting to note how closely the number of refresh time observations follow the minimum observation count for the individual stocks. Their temporal correlation is 0.82.

We use the data set to compute four different estimates of the integrated covariance matrix. The first is the realized covariance, RC_t , based on intraday returns that span a interval of 5 minutes, so there are at most 78 such returns each day (the previous-tick method is used to construct a discretized return vector). MRK_t denote the multivariate realized kernel based on refresh time synchronization of all the 473 stocks. Finally, we have the composite realized kernel, CRK_t. We compute two versions of this. One denoted CRK_t^{cor} which is exactly as described in section 2.2. The other, which we denote CRK_t^{var}, have the individual bandwidths for all the underlying 2×2 multivariate realized kernels selected as in Barndorff-Nielsen et al. (2011). That is it uses the average of optimal bandwidth from the univariate realized kernels. We will now take a closer look at characteristics of the estimates.

Figure 2: Daily average IV estimates from the three alternative estimators. Note that both composite realized kernels, CRK_t^{var} and CRK_t^{cor}, deliver precisely the same IV estimates.

In Figure 2 we display the average of the diagonal elements of RC_t, MRK_t and the 473 univariate realized kernels that appear on the diagonals of CRK_t^{cor} and CRK_t^{var} . The average estimates are very similar on a given day with univariate estimates mostly on the top, followed by MRK_t. As expected the average daily variance of the assets is highest in late 2008, with the difference between the level of daily volatility then and the more tranquil periods such November 2009 being remarkable.

In Figure 3 we present the average of absolute value of the correlations computed from RC_t, MRK_t, CRK_t^{var} and CRK_t^{cor} . We use the absolute values corresponding to the L_1-norm

time stamp.

to illustrate the overall level of dependence estimated by the alternative estimators [NOTE: Neil wrote "by L_1-norm is futher estimation error?"]. The average absolute estimates mostly have MRK_t correlation on the top, and CRK_t^{cor} on the lower side, although the dynamics are identical in all estimators. The average absolute correlation reaches high levels in March 2008 and again duraing the late 2008 crises.

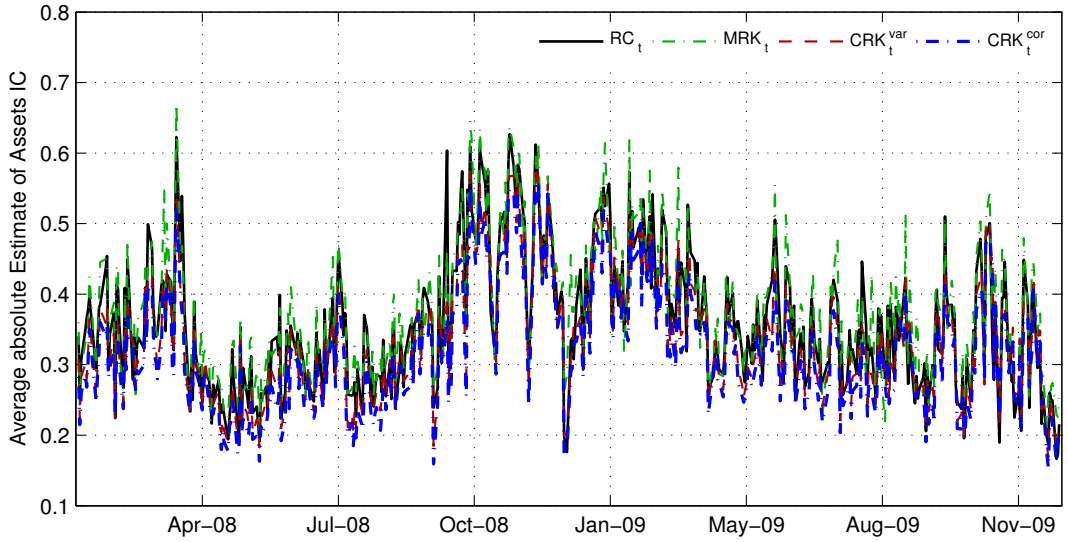

Figure 3: Daily average absolute IC estimates from the four alternative estimators.

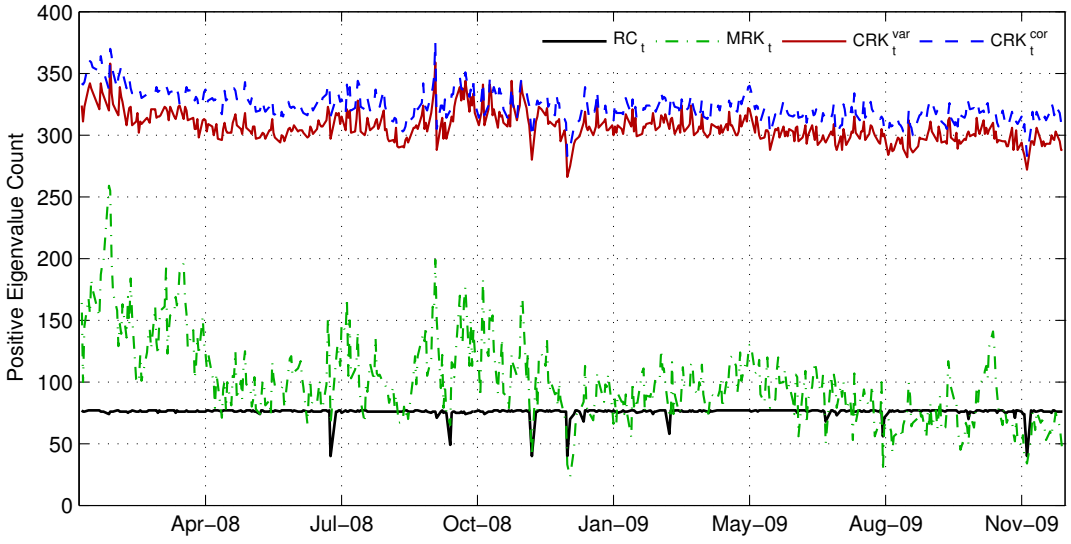

Figure 4: Daily number of positive eigenvalues.

In Figure 4 we present the daily number of positive eigenvalues of RC_t, MRK_t, CRK_t^{var} and CRK_t^{cor} . Here CRK_t^{cor} is almost always on the top, and the effect of data loss on MRK_t is clearly evident. Note that both RC and MRK have either 0 or positive eigenvalues; by construction

they cannot have negative eigenvalues. This is not the case for the raw *CRK* estimators which have negative eigenvalues on all days in the sample. Of course we may deal with this by eigenvalue cleaning later.

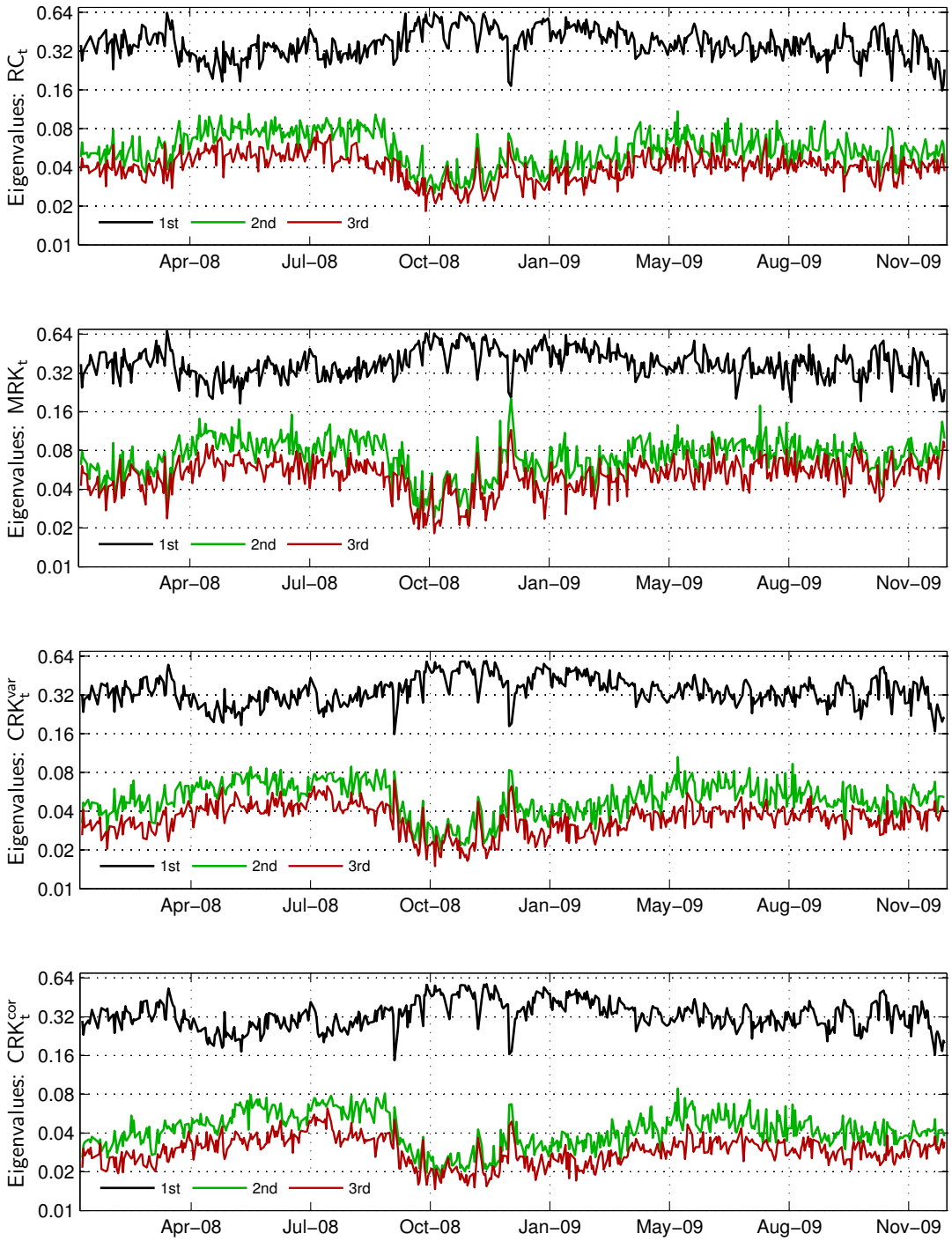

Figure 5: The first three eigenvalues for each of the four estimators [NOTE: Will rearrange to have all 1st eigenvalues at top, the 2nd below etc].

In Figure 5 we present the first three daily eigenvalues, as a ratio to the sum of all eigenvalues computed from the correlation matrices derived from RC_t, MRK_t, CRK_t^{var} and CRK_t^{cor}. These can be interpreted as a partial R^2, such that a particular eigenvalues gives how much of the total variation in the assets return is explained by the corresponding eigenvector. All estimators paint the same picture with the first eigenevalue being particlur dominant during periods of the financial crises late 2008.

4 Applications in Portfolio Construction

In this section we consider some applications in portfolio construction. We analyze the merits of our composite realized kernels in an ultra high dimensional environment. The first application is a minimum variance portfolio exercise and this is followed by an investigation of portfolio tracking.

4.1 Minimum Variance Portfolio Construction

Covariance matrix estimates are key ingredients in portfolio optimizers. The classic Markowitz (1959) portfolio problem is formulated as

$$\begin{aligned} & \min w'\Sigma w \\ \text{subject to} \quad & w'\mu = \mu \\ & w'\iota = 1 \end{aligned}$$

where w is the k by 1 vector of portfolio weights, Σ is the d by d covariance of returns, μ is the by 1 vector of expected returns, and ι is a conformable vector of 1s.

This problem is known to be sensitive to estimation of the mean returns (e.g. Jagannathan & Ma 2003), and so we focus on the simpler problem if constructing the global minimum variance portfolio (GMVP) which is the solution to

$$\begin{aligned} & \min w'\Sigma w \\ \text{subject to} \quad & w'\iota = 1. \end{aligned}$$

When Σ is positive definite, the solution to the GMVP problem is

$$w = \frac{\Sigma^{-1}\iota}{\iota'\Sigma^{-1}\iota}.$$

When the covariance is not (strictly) positive definite, as is the case for some of the estimators used in this study, the solution to this problem does not exist and so additional constraints are required. Moreover, even if the covariance is positive definite it may not be well conditioned which can produce unrealistically volatile weights. This has prompted some researchers , e.g. DeMiguel, Garlappi & Uppal (2009), to propose using simple weights such as $w_i = 1/d$ as

delivering more reliable results than when using econometric methods to estimate Σ. Recently Brodie et al. (2009), DeMiguel, Garlappi, Nogales & Uppal (2009) and Fan, Zhang & Yu (2009) have studied the problem of adding L_1constraints to the classic Markowitz portfolio problem. This modified problem is then

$$\begin{aligned} & \min w'\Sigma w \\ \text{subject to} \quad & w'\iota = 1 \\ & \textstyle\sum_{i=1}^{d} |w_i| \leq 1 + 2s, \end{aligned}$$

where s is the percentage that is allowed to be held short. When $s = 0$ the portfolio is long only. Setting $s = 0.3$ would allow for "130-30" portfolios. The L_1 constraint produces a problem which has a solution when Σ is positive semi-definite or even indefinite. A slightly modified LARS algorithm can be used to find the solution numerically (Efron, Hastie, Johnstone & Tibshirani 2004). [4,5]

4.1.1 Performance Evaluation

We use a random walk forecasting model

$$E_t[\Sigma_{t+1}] = \hat{\Sigma}_t$$

where $\hat{\Sigma}_t$ is one of the 4 estimators crossed with one of the 4 regularization methods. This is a remarkable setup, as it means that we will only use yesterday's high frequency data to perform asset allocation today. This will put extreme stress on estimating yesterday's covariance well. Of course in practise it may be wiser to use a moderate amount of time series averaging of past $\hat{\Sigma}_t$ to risk some bias but gain some estimation precision.

Portfolio weights $\hat{w}_t$ based on $\hat{\Sigma}_t$ are computed daily using as described in the previous section. Note that s gives the amount of short interest allowed in the portfolio. We will evaluate the performance of the 16 estimators for $s \in \{0\%, 25\%, 50\%, 100\%\}$, varying from a long-only portfolio to allowing for a great deal of shorting.

[4][NOTE: Move to appendix] The intuition behind this numerical strategy is simple. Let w_c be the optimal weight vector for a constraint of c. The initial set of weights, w_0, can be easily computed using a standard quadratic programming routine. This set of weights is usually sparse in the sense the many of the elements are exactly 0. The set of positive weights is known as the "active set". Let $\Sigma_{\mathcal{A}}$ be the the submatrix of Σ corresponding active set. If $\Sigma_{\mathcal{A}}$ is positive definite, for a small increase in the L_1 constraint from 0 to δ, the active set will not change, and the optimal portfolio will expand along the line containing w_0 and $\Sigma_{\mathcal{A}}^{-1}\iota / \iota\Sigma_{\mathcal{A}}^{-1}\iota$. This line balances the derivatives of the problem while respecting the constraint that the weights sum to 1. Eventually the L_1 constraint is relaxed sufficiently that a new variable enters the active set. At this point the new variable is added and the direction vector is updated. This process of moving from the current constrained optimal portfolio towards the unconstrained portfolio, adding to the active set, continues until all variables are added.

[5]If the covariance matrix is positive semi-definite, then the algorithm will end as soon as a sub-matrix is selected which has a 0 eigenvalue, and so does not apply to the covariance forecasts constructed using eigenvalue cleaning or when the covariance is projected onto a factor structure.

For a particular constraint s, our measure of performance is

$$\sigma^2_{i,t} = \hat{w}'_{i,t} RC_{t+1} \hat{w}'_{i,t} \quad i = 1, \ldots, 16.$$

Note that the portfolio-weighted 5-minute realized covariance estimate is identical to using the 5-minute realized variance of the portfolio.

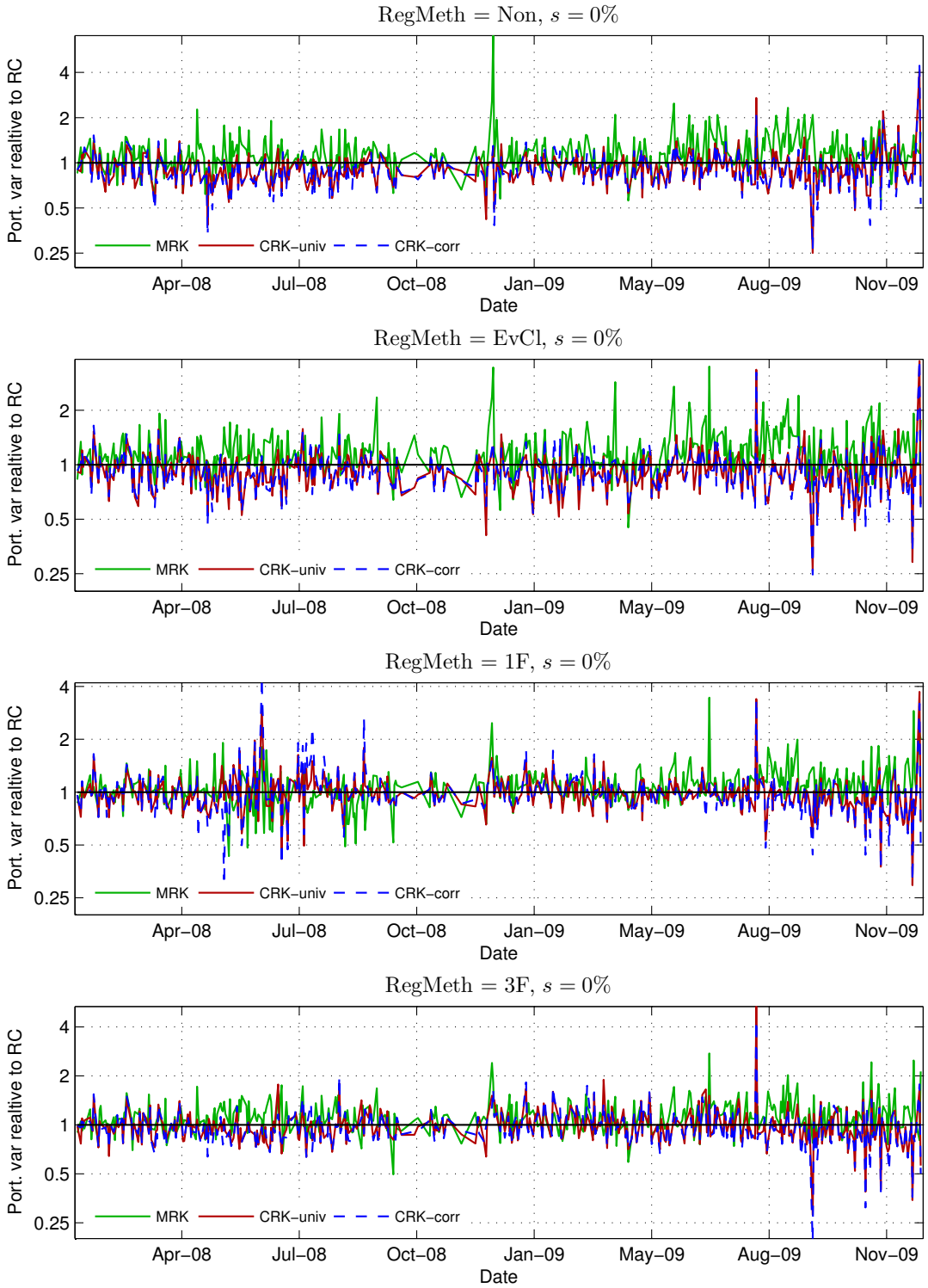

Figure 6: Ratio of the realized portfolio variance ($\sigma^2_{i,t}$) of either *MRK*, CRK^{var} or CRK^{cor} to *RC* for $s = 0\%$. Values larger than 1 indicate that the realized portfolio variance was smaller for the realized covariance.

In Figure 6 we plot the $\sigma^2_{i,t}$ series for $s = 0\%$. To enhance the visualization we have plotted MRK_t, CRK_t^{var} and CRK_t^{cor} relative to RC_t. [Add comments on results]

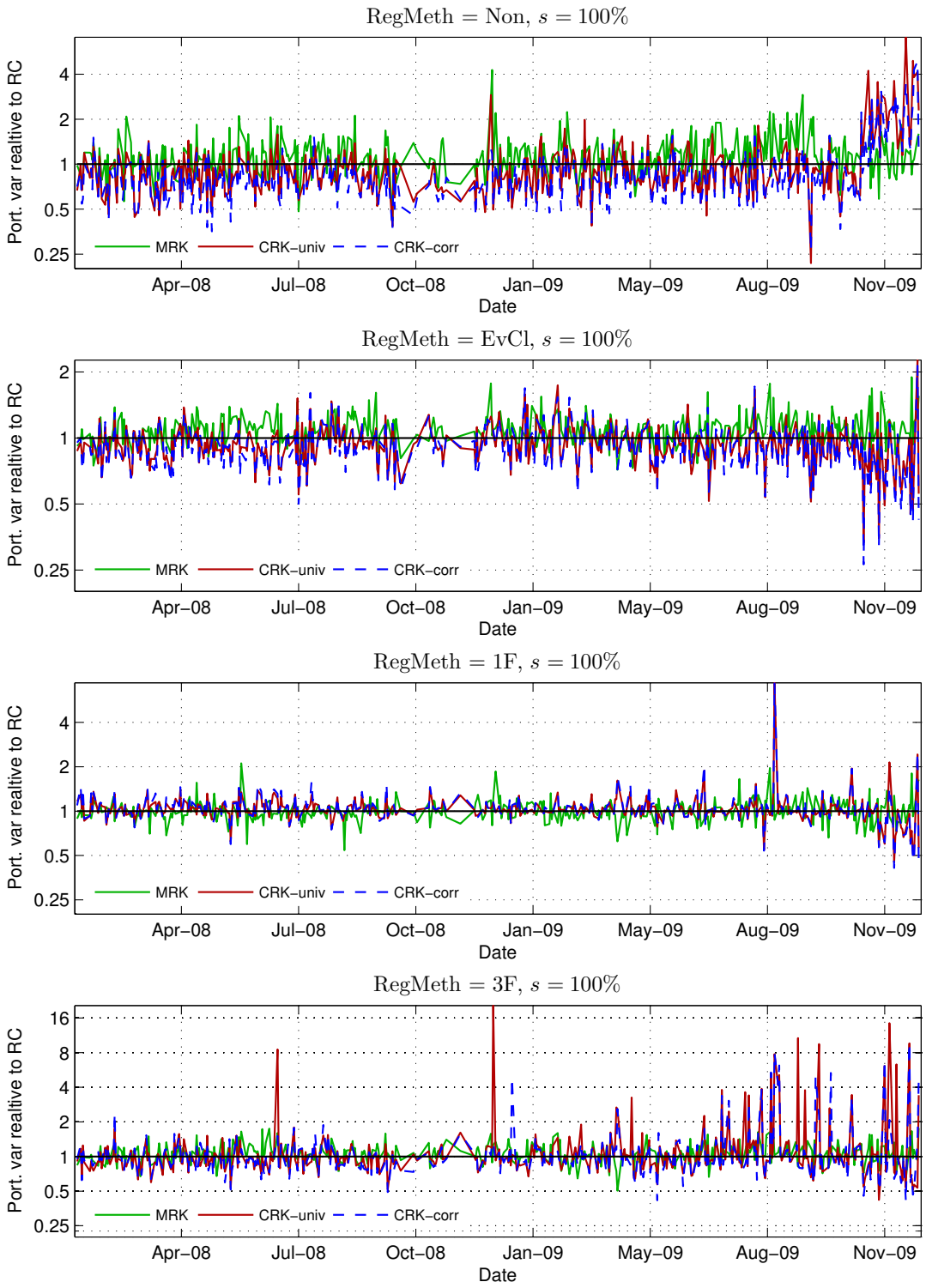

Figure 7: Ratio of the realized portfolio variance ($\sigma^2_{i,t}$) of either *MRK*, CRK^{var} or CRK^{cor} to *RC* for $s = 100\%$.. Values larger than 1 indicate that the realized portfolio variance was smaller for the realized covariance.

In Figure 7 we plot the $\sigma^2_{i,t}$ series relative to portfolio variance based on RC_t for $s = 100\%$. [Add comments on results]

4.1.2 Statistical significance in differences in performance

The average portfolio variance, $\bar{\sigma}_p^2$, for each estimator and each value of s is given in Table 1. We also include the $1/d$ portfolio amongst the alternatives. This seems to show that the $1/d$ portfolio is uncompetitive and that some short selling do reduce portfolio variance. Still, we need to assess the statistical significance of the differences.

To assess whether the differences in the selected minimum variance portfolios are statistically significant we apply the Model Confidence Set (MCS) methodology suggested in Hansen et al. (2011). The MCS is useful in a multiple comparison problems since it automatically controls the familywise error rate which is difficult in multiple pairwise comparisons.

To explain the methodology let the set $\mathcal{M}^0$ contain the alternatives that are indexed by $i = 1, \ldots, m_0$. Define the relative performance variables

$$\delta_{ij,t} \equiv \sigma_{i,t}^2 - \sigma^2{}_{j,t}, \qquad \text{for all} \qquad i, j \in \mathcal{M}^0,$$

then we rank the alternatives in terms of expected relative performance, $\mu_{ij} \equiv \mathrm{E}(\delta_{ij,t})$, so that alternative i is preferred to alternative j if $\mu_{ij} < 0$. The objective of the MCS procedure is to determine $\mathcal{M}^*$, the set of superior objects is defined by

$$\mathcal{M}^* \equiv \{i \in \mathcal{M}^0 : \mu_{ij} \leq 0 \quad \text{for all} \quad j \in \mathcal{M}^0\}.$$

This is done through a sequence of significance tests, where objects that are found to be significantly inferior to other elements of $\mathcal{M}^0$ are eliminated. So the MCS procedure yields a model confidence set, $\widehat{\mathcal{M}}^*$, that is a collection of models built to contain the best models with a given level of confidence. The set $\widehat{\mathcal{M}}^*$ includes the best model(s) with a certain probability in the same sense that a confidence interval covers a population parameter.

The MCS procedure yields p-values for each of the objects. An object with a small MCS p-value makes it unlikely that it is one of the "best" alternatives in $\mathcal{M}^0$. The MCS p-value should not be interpreted as the probability that a particular model is the best model, exactly as a classical p-value is not the probability that the null hypothesis is true. Rather, the probability interpretation of a MCS p-value is tied to the random nature of the MCS because the MCS is a *random* subset of models that contains $\mathcal{M}^*$ with a certain probability. The MCS is constructed such that inference about the "best" follows the conventional meaning of the word "significance". So a model is discarded only if it is found to be significantly inferior to another model. In our implementation of the MCS we used the Max statistics as recommended in Hansen et al. (2011). We used block length of 30 and 50,000 bootstrap re-samples.

In Table 1 we present the average portfolio variances for the alternative portfolio weights together with their MCS p-values. When the portfolios are tightly constrained ($s = 0\%$ or $s = 25\%$), the *CRK* dominate the MCS. Moreover, the gains from using the composite method are evident for all 4 methods, even when one method is not in the MCS (Eigenvalue cleaning). When the constrain is relaxed to allow for 50% short interest, the composite kernel using the

correlation weights performs the best, although the realized covariance using either the 3-Factor projection or Eigenvalue cleaning is also in the MCS. Finally, when a portfolio with 100% shorting is allowed the *CRK* is still the best performing model, although in this case using Eigenvalue cleaning. We interpret this as evidence that tight restrictions are useful when the amount of short interest is very limited, but that the tightest restrictions, that a 1-Factor model is sufficient, is overly restrictive when substantial short selling is allowed.

Table 1: Global Minimum Variance Portfolios: Performance at Full Sample.

	$s = 0\%$		25%		50%		100%	
Covariance Estimator	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}
$1/n$	1.920	0.001	1.920	0.002	1.920	0.001	1.920	0.002
Non regularized								
RC	0.749	0.004	0.633	0.009	0.648	0.002	0.675	0.002
MRK	0.814	0.001	0.697	0.001	0.697	0.001	0.701	0.001
CRK^{var}	**0.666**	**1.000**	0.512	0.017	0.488	0.013	0.509	0.002
CRK^{cor}	**0.668**	**0.790**	**0.502**	**0.333**	0.464	0.017	0.474	0.002
Eigenvalue cleaning								
RC	0.807	0.005	0.563	0.095	**0.415**	**0.262**	**0.375**	**0.488**
MRK	0.881	0.002	0.592	0.015	0.437	0.039	0.400	0.031
CRK^{var}	0.693	0.042	0.528	0.022	0.447	0.039	**0.368**	**0.488**
CRK^{cor}	0.707	0.033	0.530	0.039	0.438	0.046	**0.365**	**1.000**
1-Factor model								
RC	**0.682**	**0.423**	**0.463**	**0.986**	0.453	0.023	0.457	0.002
MRK	0.723	0.012	**0.484**	**0.613**	0.464	0.015	0.466	0.002
CRK^{var}	**0.685**	**0.147**	**0.460**	**0.986**	0.453	0.018	0.471	0.002
CRK^{cor}	**0.688**	**0.131**	**0.461**	**0.986**	0.453	0.026	0.471	0.002
3-Factor model								
RC	**0.684**	**0.206**	**0.468**	**0.971**	**0.417**	**0.234**	**0.412**	**0.137**
MRK	0.727	0.008	**0.489**	**0.662**	0.430	0.039	0.425	0.018
CRK^{var}	0.696	0.061	**0.474**	**0.865**	0.426	0.050	0.488	0.003
CRK^{cor}	**0.681**	**0.206**	**0.459**	**1.000**	**0.390**	**1.000**	0.406	0.045

This table reports the average realized variance from the 4 estimators times the 4 projection methods, and the MCS p-value. Each of the four sets of columns represents a different short selling constraint, ranging from no shorts ($s = 0\%$) to allowing for 100% short interest ($s = 100\%$). Bold entries indicate forecasts that were in the MCS using a p-value of 10%.

These results are very positive for the Composite RK, but it was somewhat of a puzzle to us as to why the RC is not completely outperformed. To see what contributed the most to these results, we identified the 5 days with the (out-of-sample) largest portfolio variances *for each estimator*. Taking the union of these days resulted in a set of 22 unique days. We deleted those days and give the "trimmed" performance results in Table 2. The performance of the composite kernel is very good in this scenario and one or both of the *CRK* are in the MCS for each of the short-sale constraints. The same pattern appears where tightly parametrized

specifications perform better when the the short-sale constraint is tight, but methods which allow for a richer structure work better when more short selling is allowed.

Table 2: Global Minimum Variance Portfolios: Performance at Trimmed Sample.

	$s = 0\%$		25%		50%		100%	
Covariance Estimator	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}
$1/n$	1.581	0.000	1.581	0.000	1.581	0.000	1.581	0.000
Non regularized								
RC	0.558	0.000	0.476	0.001	0.495	0.000	0.517	0.000
MRK	0.613	0.000	0.530	0.000	0.536	0.000	0.545	0.000
CRK^{var}	**0.494**	**0.457**	0.382	0.001	0.377	0.004	0.417	0.000
CRK^{cor}	**0.491**	**1.000**	0.369	0.008	0.356	0.029	0.384	0.000
Eigenvalue cleaning								
RC	0.570	0.002	0.400	0.003	**0.321**	**0.287**	0.301	0.000
MRK	0.629	0.000	0.426	0.001	0.342	0.042	0.323	0.000
CRK^{var}	**0.503**	**0.169**	0.382	0.002	0.335	0.078	0.283	0.000
CRK^{cor}	0.506	0.041	0.378	0.003	**0.327**	**0.261**	**0.279**	**1.000**
1-Factor model								
RC	0.514	0.041	0.368	0.008	0.372	0.001	0.377	0.000
MRK	0.546	0.000	0.384	0.001	0.381	0.000	0.385	0.000
CRK^{var}	**0.503**	**0.442**	0.363	0.031	0.370	0.001	0.385	0.000
CRK^{cor}	**0.506**	**0.308**	0.362	0.031	0.369	0.013	0.385	0.000
3-Factor model								
RC	0.507	0.074	0.361	0.038	0.339	0.047	0.339	0.000
MRK	0.546	0.001	0.382	0.002	0.355	0.006	0.354	0.000
CRK^{var}	**0.500**	**0.308**	0.355	0.038	0.330	0.078	0.368	0.000
CRK^{cor}	**0.493**	**0.779**	**0.345**	**1.000**	**0.311**	**1.000**	0.333	0.000

This table reports the average realized variance from the 4 estimators times the 4 projection methods, and the MCS p-value based on the trimmed sample. A total of 22 days were trimmed which included the 5 highest realized portfolio variances for each forecast. Each of the four sets of columns represents a different short selling constraint, ranging from no shorts ($s = 0\%$) to allowing for 100% short interest ($s = 100\%$). Bold entries indicate forecasts that were in the MCS using a p-value of 10%.

The actual portfolio variance of the trimmed days are presented in Figure 8. This shows that the difference in the realized portfolio variance was typically small in the days removed, and that most of the removed days occurred during the financial crisis of 2008. December 11, 2009, however, was a bad day for the *CRK* while much less problematic for *RC*. We were able to identify what caused the difference between the *RC* and the *CRK* on this day. Taking for example $s = 0$ (no short selling), on this day there is an extreme volatility event for the Ticker RX, and the daily estimates are as large as 66%.[6] The day before it was only 0.06%.[7] The *RC*

[6] See http://seekingalpha.com/article/178222-ims-health-oversold-on-senate-proposal .

[7] On December 10, 2010 RX traded is a very narrow range of approximately .1% of its price. The bid-ask spread was 1/3 of the daily range and so the RC estimator picked up the bid-ask bounce and over-estimated the volatility.

produces a portfolio a weight of 0.17 on this day, but *CRK* had a much larger weight of 0.43. The reason for the large load for the *CRK* is that it on Dec 10 estimated the variance of RX to be only 1/3 of what the RC estimated it to be. This suggests imposing an upper bound on the weights, e.g. $|\omega_i| \leq 10/d$.

We conducted a new set of performance measures only deleting this day. This is what is in Table 3. These results are similar to the two previous tables with a strong overall performance by the *CRK*.

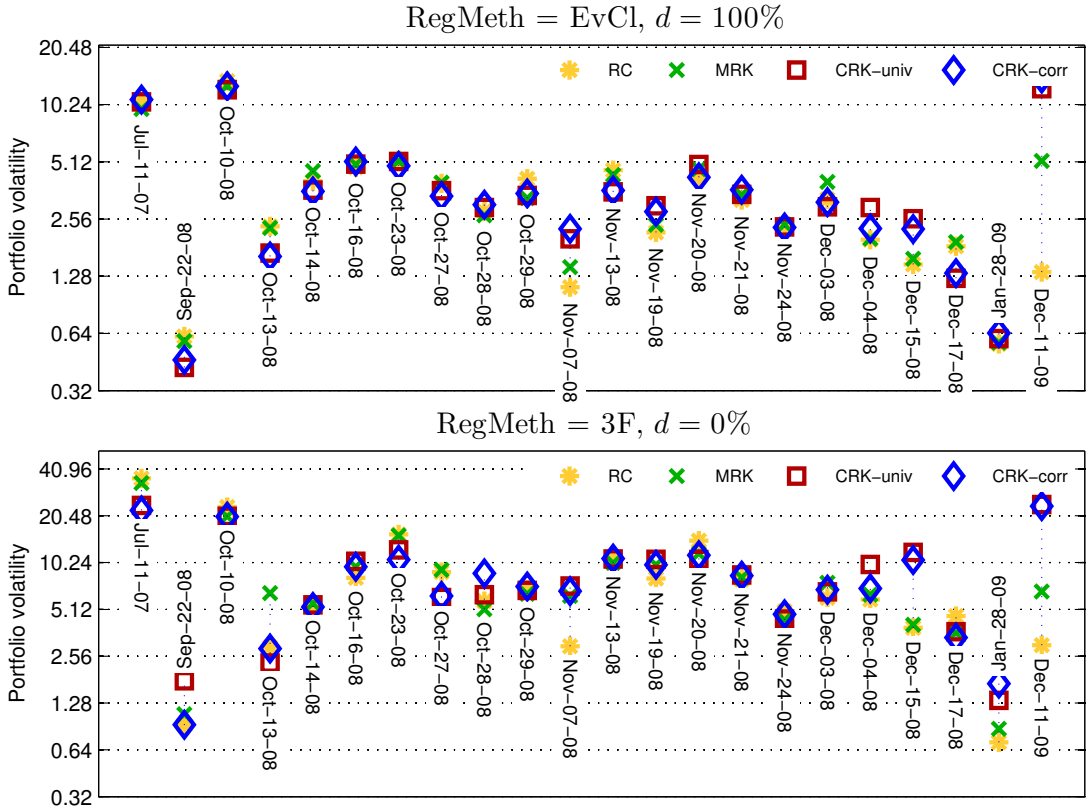

Figure 8: Portfolio variances on the trimmed days.

Table 3: Global Minimum Variance Portfolios: Only December 11, 2009 excluded.

	$s = 0\%$		25%		50%		100%	
Covariance Estimator	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}
$1/n$	1.921	0.001	1.921	0.002	1.921	0.000	1.921	0.001
Non regularized								
RC	0.748	0.003	0.632	0.005	0.647	0.003	0.674	0.001
MRK	0.800	0.002	0.683	0.001	0.683	0.000	0.687	0.001
CRK^{var}	**0.654**	**1.000**	0.502	0.006	0.482	0.007	0.507	0.001
CRK^{cor}	**0.656**	**0.851**	**0.491**	**0.129**	0.456	0.009	0.470	0.001
Eigenvalue cleaning								
RC	0.805	0.014	0.562	0.009	0.414	0.092	0.374	0.001
MRK	0.868	0.003	0.587	0.006	0.432	0.018	0.395	0.001
CRK^{var}	0.675	0.081	0.514	0.007	0.435	0.026	0.356	0.015
CRK^{cor}	0.687	0.065	0.514	0.017	0.425	0.026	**0.352**	**1.000**
1-Factor model								
RC	**0.679**	**0.201**	**0.462**	**0.844**	0.452	0.008	0.456	0.001
MRK	0.717	0.010	**0.482**	**0.116**	0.462	0.006	0.464	0.001
CRK^{var}	**0.667**	**0.517**	**0.452**	**0.849**	0.445	0.009	0.463	0.001
CRK^{cor}	**0.667**	**0.574**	**0.452**	**0.849**	0.444	0.012	0.463	0.001
3-Factor model								
RC	**0.681**	**0.111**	**0.467**	**0.532**	0.417	0.026	0.412	0.001
MRK	0.721	0.007	0.488	0.015	0.428	0.018	0.423	0.001
CRK^{var}	**0.672**	**0.347**	**0.463**	**0.600**	0.414	0.026	0.477	0.001
CRK^{cor}	**0.658**	**0.851**	**0.448**	**1.000**	**0.380**	**1.000**	0.397	0.001

This table reports the average realized variance from the 4 estimators times the 4 projection methods, and the MCS p-value based on the full sample excluding December 11, 2009. Each of the four sets of columns represents a different short selling constraint, ranging from no shorts ($s = 0\%$) to allowing for 100% short interest ($s = 100\%$). Bold entries indicate forecasts that were in the MCS using a p-value of 10%.

4.2 Tracking Portfolios

Construction of tracking portfolio is closely related problem to minimum variance construction. The problem can be formulated as

$$\begin{aligned} \min_{w} \quad & w'\Sigma w \\ \text{subject to} \quad & w_1 = 1 \end{aligned}$$

where it is assumed that the asset to be tracked is in position 1. For us assets 1 is SPY in our application. When Σ is known, the solutions is $w = [1 \ \Sigma_{22}^{-1}\Sigma_{12}]$ where the covariance matrix has been partitioned

$$\Sigma = \begin{bmatrix} \sigma_{11} & \Sigma_{12}' \\ \Sigma_{12} & \Sigma_{22} \end{bmatrix}.$$

This solution relies on Σ_{22} having full rank, and so is not necessarily applicable to large portfolios. An L_1 constrained version can be constructed as the solution to

$$\begin{aligned} \min_{w} \quad & w'\Sigma w \\ \text{subject to} \quad & w_1 = 1 \\ & \textstyle\sum_{i=2}^{d} |w_i| \leq \quad s \end{aligned}$$

where $s \in [0, \infty)$ controls the gross exposure of the tracking portfolio. The path of constrained tracking weights can be computed starting at $d = 0$ by allocating weight to the asset(s) which produces the largest decrease in the tracking error variance, where the score of this problem is

$$\frac{\partial S(w)}{\partial w} = -2\Sigma_{12} + 2\Sigma_{22}w$$

where $S(w)$ is the tracking error variance for a given weight vector w. The LARS algorithm with the LASSO modification is directly applicable to this problem, and so computing the tracking portfolio weights is a standard problem, except when the estimated covariance of the active set is not positive definite. When this occurs, we stop the weight constuction.

4.2.1 Alternative portfolio construction

An alternative is to choose a target number of assets to employ to track where the L_1constrained algorithm is used to choose the set. The is similar to applications of the LASSO in regression settings where the constrained problem is used for model selection, and then OLS is used to estimate the parameters conditional on the selected model (REFERENCE HERE). These type of solutions are using an L_1 constrained problem to approximate an L_0 problem which is to select the subset of tracking assets which produce the smallest tracking error variance subject to having no more than n assets in the tracking portfolio. The L_0 problem is computationally intractable except when the number of assets is small. We use this method for $n = 10$ and 25. The sets of assets are selected by finding the s where the 10th (or 25th) asset just enters the portfolio.

4.2.2 Results

TBC

Table 4: Global Minimum Variance Portfolios: Performance at Full Sample.

	10 assets		30 assets	
Covariance Estimator	$\overline{\sigma}_p^2$	p_{MCS}	$\overline{\sigma}_p^2$	p_{MCS}
Non regularized				
RC	0.181	0.005	0.102	0.002
MRK	0.197	0.005	0.111	0.002
CRK^{var}	**0.162**	**0.276**	**0.083**	**0.232**
CRK^{cor}	**0.160**	**1.000**	**0.082**	**1.000**
Eigenvalue cleaning				
RC	0.203	0.005	0.122	0.002
MRK	0.209	0.005	0.125	0.002
CRK^{var}	0.226	0.005	0.136	0.002
CRK^{cor}	0.228	0.005	0.138	0.002
1-Factor model				
RC	0.240	0.005	0.174	0.001
MRK	0.251	0.005	0.182	0.001
CRK^{var}	0.274	0.005	0.203	0.002
CRK^{cor}	0.269	0.005	0.199	0.002
3-Factor model				
RC	0.226	0.005	0.156	0.002
MRK	0.238	0.005	0.163	0.001
CRK^{var}	0.260	0.005	0.188	0.002
CRK^{cor}	0.256	0.005	0.183	0.002

This table reports the average realized variance from the 4 estimators times the 4 projection methods, and the MCS p-value. Each of the two sets of columns represents a different number of tracking assets. Bold entries indicate forecasts that were in the MCS using a p-value of 10%.

5 Conclusion

This paper has introduced the composite realized kernel. Composite realized kernels are data efficient estimators which make use of the maximum amount of available data to estimate each element of the integrated covariance. The individual integrated variances are estimated using all data and univariate realized kernels, and the correlations are estimated using pairwise multivariate realized kernels which minimizes the data loss through refresh time sampling. We extend the existing theoretical results to allow for optimal bandwidth selection for correlation estimation, which is generally different from what is done in the existing literature. These two innovations simplify the construction of vast dimensional covariance estimates from high-frequency data.

We compared the common 5-minute realized covariance estimator, the multivariate realized kernel, and two composite realized kernels through their ability to generate superior out-

of-sample portfolio allocations. We found that the composite realized kernel was consistently among the best performing forecasting model. We also documented that applying some form of regularization or shrinkage was useful even when the amount of short selling was constrained.

Appendix A: Random matrix theory and realised correlation

Suppose Y is a d-dimensional vector of independent scaled, by a time-invariant common σ, Brownian motions and we write

$$y_i = Y(i/n) - Y((i-1)n), \quad i = 1, 2, ..., n.$$

Then the realised covariance is

$$V = \sum_{i=1}^{n} y_i y_i'.$$

For any value of n its expectation is $\sigma^2 I_d$, so all of its eigenvalues are σ^2. Now V follows a so-called "white Wishart matrix" and its exact distribution is $W_d(\sigma^2 I_d, n)$. Of interest is the empirical distribution of the estimated eigenvalues $\widehat{\lambda}_i$ of V. We use asymptotics to study this with both n and d going to infinity. This area of study is called "random matrix theory" and surveys of the main idea and the literatures are given by Johnstone (2001) and Johnstone (2010) and a survey in the context of finance is given in Gatheral (2008) and Bouchaud & Potters (2011). Laloux et al. (2000) gives an explicit recipe for "eigenvalue cleaning" or "eigenvalue clipping" using random matrix theory, which is used by Hautsch et al. (2011). Clearly the distribution of population eigenvalues is simply a point mass on σ^2, as all the true eigenvalues are at σ^2. If we define

$$n/d \to q \geq 1$$

then the empirical distribution of the estimated eigenvalues is

$$\widehat{G}(\lambda) = \frac{1}{d} \sum_{i=1}^{d} I(\widehat{\lambda}_i \leq \lambda) \xrightarrow{u.p.} G(\lambda) = \int_{\lambda_-}^{\lambda} \rho(t) \mathrm{d}t$$

where $\rho(\lambda)$ is the Marcenko & Pastur (1967) density

$$\rho(\lambda) = \frac{q}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \quad \lambda \in (\lambda_-, \lambda_+).$$

Here the support is determined by

$$\lambda_\pm = \sigma^2 \left(1 \pm \sqrt{\frac{1}{q}}\right).$$

Hence if q is large $\lambda_{\pm} \simeq \sigma^2$, which means that the dimension of the system is sufficiently small compared to the sample size n that we can precisely estimate all the eigenvalues and so the density becomes close to have a point mass at σ^2, while if q is close to one then $\lambda_{-} \simeq 0$ and $\lambda_{+} \simeq 2\sigma^2$. Now Jiang (2004) showed the same result holds for the correlation matrix, and so applies immediately to realised correlation matrices but with $\sigma^2 = 1$.

Geman (1980) proved that if $q \geq 1$ then the largest estimated eigenvalue has the property that

$$\widehat{\lambda}_{\max} \xrightarrow{a.s.} \sigma^2 \left(1 + \sqrt{\frac{1}{q}}\right)^2.$$

So $n\widehat{\lambda}_{\max} \sim \sigma^2 \left(\sqrt{n} + \sqrt{d}\right)^2$. This holds even if $n < p$, so long as both go off to infinity (see (Johnstone 2001, p. 300)). This latter observation is important to us as we will apply it to a realised correlation with n being small compared to d but both being quite large. It will mean that λ_{+} can be well above $2\sigma^2$.

Remark 1 *This suggests a theoretically based approach is to first use a realised correlation to estimate the number of non-white eigenvalues in the data. We use this number of non-white eigenvalues to clean the realised composite kernel result in the usual way. Notice the above approach does not take into account time-varying volatility or the superior sampling properties of the composite kernel. As far as we know random matrix theory based on a stochastic volatility model is an entirely open area, but note Kondor et al. (2004) on an attempt to use it based upon a RiskMetrics type framework. Finally, it would be attractive if we could extend the realised covariance case to the realised kernel case even when there is no stochastic volatility and no noise, but where we allow the number of terms in the kernel to increase slowly with n. That is again an entirely new area.*

A.1 More information

If we wish to go beyond the probability limit, then the distribution of the largest eigenvalue of a white Wishart random correlation matrix is determined by the Tracy-Widom law of order one, whose asymptotics is given by

$$\left|\Pr\left(\frac{n\widehat{\lambda}_{\max} - \mu_{nd}}{\sigma_{nd}} < s\right) - F_1(s)\right| \leq Ce^{-cs}d^{-2/3},$$

where F_1 is a standard distribution function available in a number of packages and

$$\begin{aligned} \mu_{nd} &= \left(\sqrt{n} + \sqrt{d}\right)^2 \\ \sigma_{nd} &= \left(\sqrt{n} + \sqrt{d}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{d}}\right)^{1/3}. \end{aligned}$$

This is discussed in Johnstone (2001).

The distribution has been tabulated by Bejan (2005), with

p-value	0.995	0.975	0.95	0.05	0.025	0.01	0.005	0.001
	-4.1505	-3.5166	-3.1808	0.9793	1.4538	2.0234	2.4224	3.2724

So if we select the 0.01 level, then

$$n\widehat{\lambda}_{\max} < \mu_{nd} + 2.0234\sigma_{nd} = \left(\sqrt{n}+\sqrt{d}\right)^2 + 2.0234\left(\sqrt{n}+\sqrt{d}\right)\left(\frac{1}{\sqrt{n}}+\frac{1}{\sqrt{d}}\right)^{1/3}.$$

So

$$\widehat{\lambda}_{\max} < \left(1+\sqrt{\frac{1}{q}}\right)^2 + 2.0234n^{-1/3}\left(1+\sqrt{\frac{1}{q}}\right)(1+\sqrt{q})^{1/3}.$$

References

Aït-Sahalia, Y., Fan, J. & Xiu, D. (2010), 'High frequency covariance estimates with noisy and asynchronous financial data', *Journal of the American Statistical Association* **105**, 1504–1517.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2000), 'Great realizations', *Risk* **13**, 105–108.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2001), 'The distribution of exchange rate volatility', *Journal of the American Statistical Association* **96**, 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F. X. & Labys, P. (2003), 'Modeling and forecasting realized volatility', *Econometrica* **71**, 579–625.

Andrews, D. W. K. (1991), 'Heteroskedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica* **59**, 817–858.

Bandi, F. M. & Russell, J. R. (2008), 'Microstructure Noise, Realized Variance, and Optimal Sampling', *Review of Economic Studies* **75**, 339–369.

Bannuh, K., van Dijk, D. & Martens, M. (2009), 'Range-based covariance estimation using high-frequency data: The realized co-range', *Journal of Financial Econometrics* **7**, 341–372.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2008), 'Designing realised kernels to measure the ex-post variation of equity prices in the presence of noise', *Econometrica* **76**, 1481–1536.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2009), 'Realised kernels in practice: Trades and quotes', *Econometrics Journal* **12**, C1–C32.

Barndorff-Nielsen, O. E., Hansen, P. R., Lunde, A. & Shephard, N. (2011), 'Multivariate realised kernels: Consistent positive semi-definite estimators of the covariation of equity prices with noise and non-synchronous trading', *Journal of Econometrics* **162**, 149–169.

Barndorff-Nielsen, O. E. & Shephard, N. (2002), 'Econometric analysis of realised volatility and its use in estimating stochastic volatility models', *Journal of the Royal Statistical Society, Series B* **64**, 253–280.

Barndorff-Nielsen, O. E. & Shephard, N. (2004), 'Econometric analysis of realised covariation: high frequency covariance, regression and correlation in financial economics', *Econometrica* **72**, 885–925.

Bauer, G. H. & Vorkink, K. (2011), 'Forecasting multivariate realized stock market volatility', *Journal of Econometrics* **160**, 93–101.

Bauwens, L., Laurent, S. & Rombouts, J. V. K. (2006), 'Multivariate GARCH models: a survey', *Journal of Applied Econometrics* **21**, 79–109.

Bejan, A. I. (2005), Largest eigenvalues and sample covariance matrices. Tracy-Widom and Painleve II: Computational aspects and realization in S-Plus with applications. M.S. dissertation, Department of Statistics, The University of Warwick.

Bouchaud, J.-P. & Potters, M. (2011), Financial applications of random matrix theory: a short review, *in* G. Akemann, J. Baik & P. D. Francesco, eds, 'Handbook on Random Matrix Theory', Oxford University Press, Oxford.

Briner, B. G. & Connor, G. (2008), 'How much structure is best? A comparison of market model, factor model and unstructured equity covariance matrices', *The Journal of Risk* **10**(4), 3–30.

Brodie, J., De Molc, I. D. C., Giannoned, D. & Loris, I. (2009), 'Sparse and stable Markowitz portfolios', *Proceedings of the National Academy of Sciences* **106**, 12267–12272.

Brownlees, C. T. & Engle, R. F. (2010), Volatility, correlation and tails for systematic risk management. Unpublished paper: Stern, NYU.

Chan, L. K., Karceski, J. & Lakonishok, J. (1999), 'On portfolio optimization: Forecasting covariances and choosing the risk model', *Review of Financial Studies* **12**, 937–974.

Chiriac, R. & Voev, V. (2011), 'Modelling and forecasting multivariate realized volatility', *Journal of Applied Econometrics* . Forthcoming.

Christensen, K., Kinnebrock, S. & Podolskij, M. (2010), 'Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion model with non-synchronous trading', *Journal of Econometrics* **159**, 116–133.

Clément, E. & Gloter, A. (2011), 'Limit theorems in the fourier transform method for the estimation of multivariate volatility', *Stochastic Processes and their Applications* **121**(5), 1097 – 1124.

DeMiguel, V., Garlappi, L., Nogales, F. J. & Uppal, R. (2009), 'A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms', *Management Science* **55**, 798–812.

DeMiguel, V., Garlappi, L. & Uppal, R. (2009), 'Optimal Versus Naive Diversification: How Inefficient is the 1/N Portfolio Strategy?', *Review of Financial Studies* **22**, 1915–1953.

Efron, B., Hastie, T., Johnstone, L. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**, 407–499.

Engle, R. F. (2009), *Anticipating Correlations*, Princeton University Press.

Fan, J., Fan, Y. & Lv, J. (2008), 'High dimensional covariance matrix estimation using a factor model', *Journal of Econometrics* **147**(1), 186–197.

Fan, J., Zhang, J. & Yu, K. (2009), Asset Allocation and Risk Assessment with Gross Exposure Constraints for Vast Portfolios, Technical report, Princeton University.

Francq, C. & Zakoian, J.-M. (2010), *GARCH Models*, Wiley, Chichester.

Gatheral, J. (2008), Random matrix theory and covariance estimation. Unpublished paper: Merrill Lynch, October.

Geman, S. (1980), 'A limit theorem for the norm of random matrices', *Annals of Probability* **8**, 252–261.

Griffin, J. E. & Oomen, R. C. A. (2011), 'Covariance measurement in the presence of non-synchronous trading and market microstructure noise', *Journal of Econometrics* **160**, 58–68.

Hansen, P. R., Lunde, A. & Nason, J. M. (2011), 'The Model Confidence Set', *Econometrica* **79**, 453–497.

Hautsch, N., Kyj, L. M. & Oomen, R. C. (2011), 'A blocking and regularization approach to high dimensional realized covariance estimation', *Forthcoming in Journal of Applied Econometrics* .

Hayashi, T. & Yoshida, N. (2005), 'On covariance estimation of non-synchronously observed diffusion processes', *Bernoulli* **11**, 359–379.

Jacod, J., Li, Y., Mykland, P. A., Podolskij, M. & Vetter, M. (2009), 'Microstructure noise in the continuous case: the pre-averaging approach', *Stochastic Processes and Their Applications* **119**, 2249–2276.

Jagannathan, R. & Ma, T. (2003), 'Risk reduction in large portfolios: Why imposing the wrong constraints helps', *Journal of Finance* **58**(5), 1651–1683.

Jiang, T. (2004), 'The limiting distributions of eigenvalues of sample correlation matrices', *Sankhya* **66**, 35–48.

Jin, X. & Maheu, J. M. (2010), Modelling realized covariances and returns. Working paper: Department of Economics, University of Toronto.

Johnstone, I. M. (2001), 'On the distribution of the largest eigenvalue in principal components analysis', *Annals of Statistics* **29**, 295–327.

Johnstone, I. M. (2010), Largest eigenvalues and eigenvectors in multivariate analysis. Unpublished: New England Statistics Symposium lecture, April.

Kondor, I., Pafka, S. & Potters, M. (2004), Exponential weighting and random-matrix-theory-based filtering of financial covariance. unpublished paper, arXiv: cond-mat/0402573.

Laloux, L., Cizeau, P., Bouchaud, J.-P. & Potters, M. (2000), 'Random matrix theory and financial correlations', *International Journal of Theoretical and Applied Finance* **3**, 391–397.

Lindsay, B. (1988), Composite likelihood methods, *in* N. U. Prabhu, ed., 'Statistical Inference from Stochastic Processes', Amercian Mathematical Society, Providence, RI, pp. 221–239.

Malliavin, P. & Mancino, M. E. (2002), 'Fourier series method for measurement of multivariate volatilities', *Finance and Stochastics* **6**, 49–61.

Malliavin, P. & Mancino, M. E. (2009), 'A fourier transform method for nonparametric estimation of multivariate volatility', *Annals of Statistics* **37**, 1983–2010.

Marcenko, V. & Pastur, L. A. (1967), 'Distribution of eigenvalues for some sets of random matrices', *Sbornik: Mathematics* **1**, 457–483.

Markowitz, H. (1959), *Portfolio Selection: Efficient Diversification of Investments*, John Wiley.

Newey, W. K. & West, K. D. (1987), 'A simple, positive semi-definite heteroskedasticity and autocorrelation consistent variance covariance matrix', *Econometrica* **55**, 703–708.

Noureldin, D., Shephard, N. & Sheppard, K. (2011), 'Multivariate high-frequency-based volatility (heavy) models', *Journal of Applied Econometrics* . Forthcoming.

Renò, R. (2003), 'A closer look at the epps effect', *International Journal of Theoretical and Applied Finance* **6**, 87–102.

Silvennoinen, A. & Teräsvirta, T. (2009), Multivariate GARCH models, *in* T. G. Andersen, R. A. Davis, J. P. Kreiss & T. Mikosch, eds, 'Handbook of Financial Time Series', Springer-Verlag, pp. 201–229.

Varin, C. (2008), 'On composite marginal likelihoods', *Advances in Statistical Analysis* **92**, 1–28.

Varin, C., Reid, N. & Firth, D. (2011), 'An overview of composite likelihood methods', *Statistica Sinica* **21**, 5–42.

Voev, V. (2008), Dynamic modelling of large-dimensional covariance matrices, *in* L. Bauwens, W. Pohlmeier & D. Veredas, eds, 'High Frequency Financial Econometrics', Physica-Verlag HD, pp. 293–312.

Voev, V. & Lunde, A. (2007), 'Integrated covariance estimation using high-frequency data in the presence of noise', *Journal of Financial Econometrics* **5**, 68–104.

Wang, Y. Z. & Zou, J. (2010), 'Vast volatility matrix estimation for high-frequency financial data', *Annals of Statistics* **38**, 943–978.

Xiu, D. (2010), 'Quasi-maximum likelihood estimation of volatility with high frequency data', *Journal of Econometrics* **159**, 235–250.

Zhang, L. (2006), 'Efficient estimation of stochastic volatility using noisy observations: A multi-scale approach', *Bernoulli* **12**, 1019–1043.

Zhang, L. (2011), 'Estimating covariation: Epps effect and microstructure noise', *Journal of Econometrics* **160**, 33–47.

Zhang, L., Mykland, P. A. & Aït-Sahalia, Y. (2005), 'A tale of two time scales: determining integrated volatility with noisy high-frequency data', *Journal of the American Statistical Association* **100**, 1394–1411.